A Prominent Solution to Test Academic and Scientific Integrity using LSA

S. Anita, M. Banu, M. N. Nachappa

Abstract: ***"Taking over the ideas, methods, or written words of another, without acknowledgment and with the intention that they be taken as the work of the deceiver" is a quotation defined by American Association of University Professors in 1989 for Plagiarism. As the above quotation states, plagiarism has been traditionally defined as the taking of words, images, ideas, etc. from an author and presenting them as one's own. It is often associated with phrases, such as capturing of words, ideas and literary theft. Plagiarism can manifest itself in a variety of ways and it is not just confined to student papers or published articles or books. For example, consider a scientist, who makes a presentation at a conference and discusses at length an idea or concept that had already been proposed by someone else and that is not considered common knowledge. During his presentation, he fails to fully acknowledge the specific source of the idea and, consequently, misleads the audience into thinking that he was the originator of that idea. This, too, may constitute an instance of plagiarism. A small number of students, about 10%, admit that they started plagiarizing because of the Internet [3]. This research studies about the concept of plagiarism with respect to internet to find the originality of a student or an author, who made a publication. It also proposes an idiosyncratic tool for identifying plagiarism of a key document by comparing the similarity of the key document with the documents in the internet pool and the results will be provided in terms of similarity percentage. This tool will be used to decide the integrity of a student or an author, who published an article by scanning through the documents available in the web.***

Index terms: Plagiarism, Idea Plagiarism, Self Plagiarism, Academic and Scientific Integrity, Latent Semantic Analysis.

I. INTRODUCTION

Scientific writing can be a complex and arduous process, for it simultaneously demands clarity and conciseness; two elements that often clash with each other. In addition, accuracy and integrity are fundamental components of the scientific enterprise and, therefore, of scientific writing. Thus, good scientific writing must be characterized by clear expression, conciseness, accuracy of what is being reported, and perhaps most importantly, honesty. Unfortunately, writing, or for that matter the entire scientific process, often occurs within the constraints of tight deadlines and other competing pressures. As a result of these constraints, scientific papers, whether generated by science students or by seasoned professionals, will at times be deficient in one or more of the above components. Insufficient clarity or lack of conciseness is typically unintentional and relatively easy to remedy by standard educational or editorial steps. Lapses in the accuracy of what is reported (e.g., faulty observations, incorrect interpretation of results) are also assumed to be most often unintentional in nature, but such lapses, even if unintentional, can have significant undesirable consequences if not corrected. Intentional lapses in integrity, even if seemingly minor, are by far the most serious type of problem because such misconduct runs contrary to the primary goal of the scientific enterprise, which is the search for truth. In scientific writing, perhaps the most widely recognized unethical lapse is plagiarism. Plagiarism can occur in many forms and some of the more subtle instances, while arguably unethical in nature, may not be classified as scientific misconduct by federal agencies such as the National Science Foundation (NSF) or the Office of Research Integrity (ORI). Nevertheless, the ethical professional is expected to operate at the highest levels of scientific integrity and, therefore, must avoid all forms of writing that could be conceptualized as plagiarism. There are other questionable writing practices, some of which may be quite common in professional scientific writing. One example is reporting and discussing results of one's research in the context of literature that is supportive of our conclusions while at the same time ignoring evidence that is contrary to our findings. Another writing 'malpractice' occurs when another author's review of a literature is used, yet the reader is led to believe that the current author has conducted the actual review. Universities throughout the world have become concerned with the question of how to minimize and respond appropriately to student plagiarism and other forms of cheating. Australian universities are highly active in educating students about plagiarism and in detecting breaches of their academic expectations.

II. CURRENT STATE OF ART

According to Carroll [6], it doesn't seem that there has been any overwhelming increase in plagiarism because of the Internet, but "it does appear that students who were plagiarizing from written sources have switched their allegiance because of the Internet. A small number of students, about 10%, admit that they started plagiarizing because of the Internet." Oliphant [3,4,5] believe that technology *has* worsened the problem. "It's never been as easy and prevalent," said Murray [2]. "For kids it's become an 'us vs. them' game, and that's not the purpose of education. Some students have developed the belief that the purpose of

Manuscript received March 22 , 2013

S. Anita, Asst. Professor, Dept. of Computer Science, St. Josephs College (Autonomous), Bangalore, Karnataka, INDIA.
M. Banu, Asst. Professor, Dept. of Computer Science, St. Josephs College (Autonomous), Bangalore, Karnataka, INDIA.
M.N. Nachappa, Assoc. Professor, Dept. of Computer Science, St. Josephs College (Autonomous), Bangalore, Karnataka, INDIA.

schooling is not the gaining of knowledge, but how I can get the degree with the least amount of work." According to Lathrop and Foss, students vary in their cut- and paste techniques [12]. Some students download and print the paper, create a title page, and hand it in [12]. The more sophisticated [students] *massage* the text, perhaps using a thesaurus to replace words or phrases the teacher might recognize as beyond their usual vocabulary or writing style. Obvious strings of highly distinctive words can be changed or deleted if a student knows the teacher is Internet-savvy and might search for strings of words online. Students find it easy to rationalize cheating. They cite unrealistic parent demands, competition for college and class rank, fear of failure, poor time management skills, sports eligibility, and time constraints compounded by after- school jobs and extracurricular activities [6]. Additionally, many feel the risk of getting caught is extremely low. Their teachers are not knowledgeable enough or familiar enough with the online turf to catch them. Some think that their teachers really don't care enough to pursue a suspicion of plagiarism. In Davis' research, many students responded that they believe high school is a joke, something they have to get through [7]. They'll get to truly meaningful work when they get into college. Students complained of boring assignments and the stress of their after-school jobs.

III. PLAGIARISM AND ITS TYPES

Plagiarism includes the following:

- Copying or paraphrasing from books or other sources without citing it properly.
- Copying work from another student.
- Working as a group on projects where the instructor requires individual work.
- Buying or copying entire Papers or projects done by others.
- Altering information or data.
- Using misleading references.

Resubmitting previously evaluated work of our own without the consent of our current instructor (e.g., submitting work, even if you have revised it, that you have previously submitted in a different course).

Work Plagiarism

Plagiarism for the purpose of this Policy and Procedure (which applies to students enrolled in course work degrees) means presenting another person's Work as one's own Work by presenting, copying or reproducing it without Acknowledgement of the Source. Plagiarism includes presenting Work for assessment, publication, or otherwise, that includes sentences, paragraphs or longer extracts from published or unpublished Work (including from the Internet) without Acknowledgement of the Source; or the Work of another person, without Acknowledgement of the Source and presented in a way that exceeds the boundaries of Legitimate Cooperation. Plagiarism can be negligent (Negligent Plagiarism) or dishonest (Dishonest Plagiarism).

a. Negligent Plagiarism

Negligent Plagiarism means innocently, recklessly or carelessly presenting another person's Work as one's own Work without Acknowledgement of the Source. Negligent Plagiarism often arises from a student's fear of paraphrasing or writing in their own words, and/or ignorance of this Policy and Procedure. It arises from failure to follow appropriate referencing practices and failure to determine or verify and acknowledge the source of the Work.

b. Dishonest Plagiarism

Dishonest Plagiarism means knowingly presenting another person's Work as one's own Work without Acknowledgement of the Source. Alleged Plagiarism will be deemed to be alleged Dishonest Plagiarism where substantial proportions of a student's work have been copied from the Work of another person, in a manner that clearly exceeds the boundaries of Legitimate Cooperation; a student's Work contains a substantial body of copied material (including from the Internet) without Acknowledgement of the Source, and in a manner that cannot be explained as Negligent Plagiarism; there is evidence that the student engaged another person to produce or conduct research for the Work, either partly or wholly, for payment or other consideration; or the student has previously received a Written Warning.

Idea Plagiarism

Appropriating an idea (e.g., an explanation, a theory, a conclusion, a hypothesis or a metaphor) in whole or in part, or with superficial modifications without giving credit to its originator. In the sciences, as in most other scholarly endeavors, ethical writing demands that ideas, data, and conclusions that are borrowed from others and used as the foundation of one's own contributions to the literature, must be properly acknowledged. The specific manner in which we make such acknowledgement varies from discipline to discipline. However, source attribution typically takes the form of either a footnote or a reference citation.

Self-Plagiarism

When plagiarism is conceptualized as theft, the notion of self-plagiarism may seem impossible. After all, one might ask: Is it possible to steal from oneself? As Larry [9] points out, it is possible to steal from oneself as when one engages in embezzlement or insurance fraud. In writing, self-plagiarism occurs when authors reuse their own previously written work or data in a 'new' written product without letting the reader know that this material has appeared elsewhere. According to Schein [11], "the essence of self plagiarism is [that] the author attempts to deceive the reader".

Internet Plagiarism

The internet can be a great source of information and an effective research tool. However, just because electronic information is easily available does not mean it is 'free'. Remember that the information you find online should be referenced, just like any other source. Online sources should be used with

care, fully acknowledged and evaluated in the same way you would any print-based source of information. Some of the common forms of Internet Plagiarism are Downloading an assignment from an online source and submitting it as our own work, Buying, stealing or borrowing an assignment and submitting it as our own work, Copying, cutting and pasting text from an electronic source and submitting it as our own work, Using the words of someone else and presenting them as our own. Copying a section of a book or an article and submitting it as our own work (that is, without acknowledgement) is plagiarism. Using significant ideas from someone else and presenting them as our own and putting someone else's ideas into our own words and not acknowledging the source of the ideas is plagiarism. Copying the written expressions of someone else without proper acknowledgment, Quoting from a source 'word for word', without using quotation marks is plagiarism, Lifting sentences or paragraphs from someone else, even with proper acknowledgment, gives the impression that the idea or information comes from the source cited, but that the phrasing, the choice of words to express it, is our own contribution. Relying too much on other people's material, repeated use of long quotations, too many direct quotations (even with quotation marks and with proper acknowledgment) result in our sources speaking, meaning our own contribution is minimal.

IV. THE PROPOSED IDIOSYNCRATIC TOOL FOR IDENTIFYING ACADEMIC AND SCIENTIFIC INTEGRITY

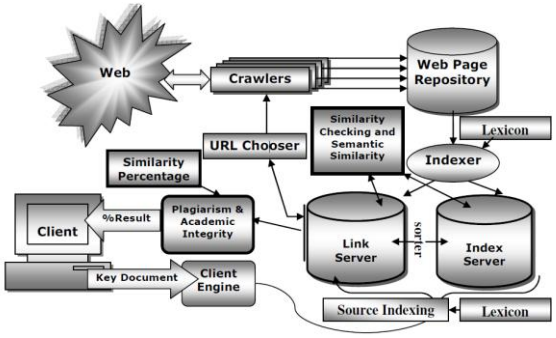

Fig 1. An Idiosyncratic Search Tool for identifying Internet Plagiarism

The proposed idiosyncratic search tool for identifying Internet Plagiarism as in Fig 1 works similar to a search engine.

The Proposed Design of Work

A conventional search engine uses a key index term for searching documents in the web. But this search tool uses a key document file consisting of several indices as the key for checking the similarity of documents in the web. First the documents available in various web servers of the web are downloaded by the crawler using either depth first crawler or breadth first crawler and the document corpus collected are stored in the web page repository i.e., a web warehouse. Both the document corpus in the web repository and the key document from the client engine are indexed using forward and inverse indexing methods. The URL addresses and the structural connectivity of the web documents are stored in the link server and the indices with respect to their document ID are stored in the index server. During indexing, a lexicon (or English Word Dictionary) has been used to mark the documents with their document id and the indices with their indices id. The key document Indices are cross-verified with the web document indices for similarity checking. The similarity percentage is calculated based upon the HIT values of the indices and the Hub/Authority Scores of the hyper lingual patterns of the Documents.

Latent semantic analysis (LSA) is used for analyzing relationships between a set of documents. Singular value decomposition (SVD) is used to preserve the similarity structure among rows. Words are then compared by taking the cosine of the angle between the two vectors formed by rows. LSA Document Similarity Checking algorithm has been used here to find out the similarity percentage of the documents in the search result. If no search results, it indicates the document's originality. The severity of Internet plagiarism or integrity of the author is identified using the Overall similarity percentage of the key document. More the number of search results for a key document, more will be the internet plagiarism and less will be the integrity of the author.

Result and Analysis

The following Table 1 represents some of the observations and findings derived from the search results. Precision and Recall values for the above search tool has also been discussed in Table 2 later.

Table 1. Observations and Findings

OBSERVATIONS	FINDINGS
No of pages crawled/second	~9 Web pages
Total pages crawled	216 Web pages
No of similarities Tested	9 key pages
Total websites visited	42 Web sites
Total no of similar pages	59 pages
Total no of dissimilar pages	155 pages
Total similar indices parsed	114 tokens
Time taken for crawling	10.2 Seconds
Time taken for indexing	5.8 Seconds
Time taken for sorting and similarity checking	5.5 Seconds

A large collection of document corpora is divided into passages with coherent meanings, typically paragraphs or documents.

The collected document corpora are then represented as a term-passage matrix. Values close to 1 represent very similar words while values close to 0 represent very dissimilar words. The calculations are based on the comparison of search results with Google/Altavista. The following Table 2 represents the recall, precision values and overall similarity percentage of search results of the

idiosyncratic tool. The memory could not be identified easily, as the size of the document corpus varies. From the above results, it has been found that all the key documents tested are having some similarity percentage. If it crosses >= a threshold value (>=60%), it has been found to be marked as Plagiarism.

Table 2. Precision (P), Recall (R) & Similarity (S)

Key Documents	w	x	y	P	R	S
Doc1.html	18	72	0	1.00	0.2	89%
Doc2.html	13	19	0	1.00	0.4	78%
Doc3.html	12	18	4	0.75	0.4	62%
Doc4.html	24	16	10	0.71	0.6	45%
Doc5.html	31	21	14	0.69	0.6	53%
Doc6.html	17	4	9	0.65	0.8	64%
Doc7.html	90	22	56	0.62	0.8	54%
Doc8.html	48	12	31	0.54	0.8	35%
Doc9.html	72	18	84	0.46	0.8	27%

A search engine's quality is measured in terms of Precision (P) and Recall (R). Relevancy is a major factor in measuring this precision and recall values. The total number of relevant documents retrieved from the search results of a particular document index is called as relevancy. If 'w' is no. of retrieved relevant, 'x' is no. of non-retrieved relevant, and 'y' is no. of retrieved irrelevant, then precision is calculated as P = w / (w+y) and recall is calculated as R = w / (w+x).

V. RELATED WORKS

Plagiarism is the practice of intentionally or unintentionally using someone else's intellectual property without properly acknowledging the original source [1]. It includes, but is not limited to, the appropriation of, buying, receiving as a gift, or obtaining by any means material that is attributable in whole or in part to another source, including words, ideas, illustrations, structure, computer code, and other expression or media, and presenting that material as one's own academic work being offered for credit [2]. Plagiarism involves submitting the same assignments in two or more classes; and using another author's ideas and argumentative forms, direct quotations, phrases and unique terminology without proper attribution. Moreover, plagiarism involves paraphrasing and summarizing without using proper attribution. Intentional, or accidental, plagiarism is perceived as a specific form of cheating, which usually occurs when a student is working independently on an assignment. University of Southern California defines plagiarism as follows [7]: it is the deliberate act of taking and using another person's work as our own. It includes absent references, reproducing the work (even with small changes) of another, taken from books, journals, articles, TV programmes, the Internet, lecture notes and so on. It also includes self plagiarism, i.e. submitting own work for more than one assessment, copying another person's work, with or without his/her consent. Also included is collusion where a group of people collaborate or collude to present an assessment or a substantial part thereof, when the examiner required individual research and outcome. According to Paul Gray [10], authors of *Student Cheating and Plagiarism in the Internet Era: A Wake-Up Call* [9], "Cheating and plagiarizing appear to be so widely accepted by students that the byword has changed from *Don't cheat or plagiarize* to *Don't get caught."* In a study of nearly 4,500 high school students, Diane Carroll [8] found the following:

- 74% of respondents reported one or more instances of serious test-cheating.
- 72% reported one or more instances of serious cheating on written work.
- 52% of students admitted they had engaged in some level of plagiarism on written assignments using the Internet.

VI. CONCLUSION

This research noted the pervasive attitude among the students documents surveyed: If it's on the Internet, its public information and we don't need to cite it. We are raising a generation of students who think anything on the Internet is free. The Internet is so anonymous and pervasive; students believe they are simply using the resources available to them. The problem of plagiarism extends well beyond the student

accused of cheating. Students who choose honesty are serious victims of this culture. The sad fact is that cheating is widespread in our culture. But, a growing truth is that sometimes cheaters *do* get caught and that dishonest behavior is getting press. Plagiarism is wrong because of the following reasons:

- Plagiarism deprives the original creator of the recognition he or she deserves.
- Plagiarism improperly allows the plagiarizer to take credit for words or ideas that he or she did not develop.
- Plagiarism is unfair to other students who diligently exert their own efforts to create a quality piece of work, which may be compared with the scholarly, yet plagiarized work of another.
- Plagiarism prevents the plagiarizer from learning and developing his or her own ideas.
- Plagiarism is not accepted in other fields such as business, science, etc.

ACKNOWLEDGMENT

A work of this nature requires the blessings and patience of a number of people. We would like to acknowledge the guidance of all our colleagues in this work. We would also like to thank god for his blessings and acknowledge his presence in all our efforts.

REFERENCES

1. Leland, B. (2000). Plagiarism and the Web. Macomb, IL: Western Illinois University. Retrieved October 31, 2000: ttp://www.wiu.edu/ users/ mfbhl/ wiu/ plagiarism.htm
2. Murray, B. (2002). Keeping plagiarism at bay in the Internet age. Monitor on Psychology, 22-24.
3. Oliphant, T. (2002). Detecting plagiarism. Edmonton, Alberta: University of Alberta. Retrieved May 15, 2002: http://www.library.ualberta.ca/ guides/ plagiarism/ detecting.html
4. Oliphant, T. (2001a). Preventing plagiarism. Edmonton, Alberta: University of Alberta. Retrieved May 15, 2002:

http://www.library.ualberta.ca/ guides/ plagiarism/ preventing.html
5. Oliphant, T. (2001b)Why students plagiarize. Edmonton, Alberta: University of Alberta. Retrieved May 15, 2002: http://www.library.ualberta.ca/ guides/ plagiarism/ students.html
6. Carroll, J., (2002), A Handbook for Deterring Plagiarism in Higher Education, Oxford Centre for Staff and Learning Development, Oxford.
7. Davis, U. C., (2001), University of Southern California, Avoiding Plagiarism: Mastering the Art of Scholarship <http://sja.ucdavis.edu/avoid.htm>
8. Diane Carroll. (2002). Teacher Quits In Dispute with School Board Over Student Plagiarism. Kansas City Star, p. 1.
9. Larry J. Sabato. (2002). Joseph Biden's Plagiarism; Michael Dukakis's 'Attack Video' – 1988. http:/www.washingtonpost.com //wp-specialreports/ clinton/ frenzy/ biden.htm
10. Paul Gray. (2002). Other People's Words. Smithsonian Magazine. http:/www.smithsonianmag.si.edu/smithsonian/issues02/mar02/presence.html
11. Schein, M. (2001). Redundant publications: From self-plagiarism to "Salami-Slicing". New Surgery, 1, 139-140.
12. Standler, R. B. (2003). Plagiarism in Colleges in USA. Retrieved February 17th, 2003 from http://www.rbs2.com/plag.htm.
13. Susan T. Dumais (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology 38: 188. doi:10.1002/aris.1440380105.